12-15-2015

Multi-Criteria Evaluation of CMIP5 GCMs for Climate Change Impact Analysis

Ali Ahmadalipour
Portland State University

Arun Rana
Portland State University

Hamid Moradkhani
Portland State University, hamidm@pdx.edu

Ashish Sharma
University of New South Wales

Multi-Criteria evaluation of CMIP5 GCMs for Climate Change Impact Analysis

Ali Ahmadalipour[1*], Arun Rana[1], Hamid Moradkhani[1], and Ashish Sharma[2]

[1]Department of Civil and Environmental Engineering, Portland State University

Portland, OR, 97201

[2]University of New South Wales, Sydney, Australia

*Corresponding Author. Email: aahmad2@pdx.edu

Accepted for publication in Theoretical and Applied Climatology

November 28, 2015

Abstract

Climate change is expected to have severe impacts on global hydrological cycle along with food-water-energy nexus. Currently, there are many climate models used in predicting important climatic variables. Though there have been advances in the field, there are still many problems to be resolved related to reliability, uncertainty and computing needs, among many others. In the present work, we have analyzed performance of 20 different Global Climate Models (GCMs) from Climate Model Intercomparison project Phase 5 (CMIP5) dataset over the Columbia River Basin (CRB) in the Pacific North-West USA. We demonstrate a statistical multi-criteria approach, using univariate and multivariate techniques, for selecting suitable GCMs to be used for climate change impact analysis in the region. Univariate methods includes Mean, Standard deviation, Coefficient of Variation, Relative Change (Variability), Mann-Kendall Test, and Kolmogorov-Smirnov test (KS-test); whereas multivariate methods used were Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Canonical Correlation Analysis (CCA), and Cluster Analysis. The analysis is performed on raw GCM data, i.e. before bias correction, for precipitation and temperature climatic variables for all the 20 models to capture the reliability and nature of particular model at regional scale. The analysis is based on spatially averaged datasets of GCMs and observation for the period of 1970 to 2000. Ranking is provided to each of the GCMs based on the performance evaluated against gridded observational data on various temporal scales (daily, monthly, and seasonal). Results have provided insight into each of the methods and various statistical properties addressed by them employed in ranking GCMs. Further; evaluation was also performed for raw GCM simulations against different set of gridded observational dataset in the area.

Keywords: Statistical Multi-Criteria Analysis, Climate Change, Pacific North-West (PNW), Columbia River Basin, Global Climate Model (GCM)

Introduction

Climate change is affecting environmental systems at global and regional scales (Moradkhani et al. 2010; Woldemeskel et al. 2012; Wang et al. 2013; Önol et al. 2014). Over the past decades, several institutions have provided future climate datasets for the Intergovernmental Panel on Climate Change (IPCC) (Pierce et al. 2009; Rupp et al. 2013), which in turn have been widely used to study climate change impacts. The World Climate Research Programme's Coupled Model Inter-comparison Project Phase 5 (CMIP5) is the latest dataset available. There have been significant improvements from the former counterparts, in knowledge and understanding of the climate using these new generation climate models. Despite these improvements, there are still large uncertainties associated with the climatic scenarios. Reliability of GCMs to simulate observed climate and consequently climatic scenarios at a regional scale is still of major concern (Rupp et al. 2013).

Evaluation of uncertainties associated with GCMs is an important aspect to consider when assessing future scenarios, e.g. their capability to simulate reliable fine scale datasets. It has been widely discussed and accepted that model uncertainty plays a big role in future projections of climatic data (Hawkins and Sutton 2011; Najafi et al. 2011). The estimation of model efficiencies is based on their performance under current or past climate conditions, and to some extent requires extrapolation to future conditions; although there are reported issues with the assumption of stationarity (Buser et al. 2009; Christensen et al. 2010). The large number of datasets, offered by various scenarios/forcings and models, adds to the uncertainty to be dealt with, along with the huge computational needs, among other varied concerns. It also adds to the ongoing debate about the reliability of GCMs to resolve features at local scale, which often are downscaled using statistical or dynamical downscaling techniques (Fowler et al. 2007; Samadi et al. 2013). Understanding of

the model processes would provide more reliable results and thus reliable future predictions. Since GCMs produce results on global scale (coarser resolution, table 1), they tend to over/under-estimate climatic variables on regional and global scales, failing to resolve the micro-scale climate. Furthermore, due to natural variability in GCM predictions, uncertainty is inevitable in their predictions. The natural variability of GCMs is higher in finer temporal scales, and thus predictions at various timescales reveal different uncertainties (Hawkins and Sutton, 2011). Therefore, it is necessary to study GCMs at different regions and assess their performance in predicting/replicating the observed climate of the region, which would further reduce the computational needs and decrease the uncertainty associated with climate prediction. Each model accounts for large amounts of climatological information leading to huge data size which in turn requires vast computations. Thus, selecting models that aptly represents the regional scale climate is a necessary first step before a regional climate change impact assessment can be performed.

Researches have been conducted in the past decade with the intention of providing ranking to GCMs performance with varied intents. Both qualitative and quantitative methods have been suggested in literature (Maxino et al. 2008; Pincus et al. 2008; Chiew et al. 2009; Christensen et al. 2010; Johnson et al. 2011) Miao et al. (2012) used four metrics, along with fitting Probability density functions (PDFs), to analyze the performance of CMIP3 precipitation and temperature datasets for China in historical period of 1960 to 1999. Rana et al. (2013) analyzed a five model ensemble of daily observed precipitation series over the period of 1961 to 2009 for Gothenburg, and assessed each model's performance. They used statistical analysis for daily and multi-day extremes, among others. Wójcik (2014) evaluated variability of GCMs in 45 CMIP5 GCMs over Europe and North Atlantic. Basic statistical methods of MAE (Mean Absolute Error), correlation coefficient, and standard deviation were used to assess the reliability of GCMs in reproducing

atmospheric circulation patterns in historical period of 1971-2000. Raju and Nagesh Kumar (2014) ranked 11 GCMs over India for the climate variable 'precipitation rate' using 5 performance indicators (correlation coefficient, normalized root mean square error, absolute normalized mean bias error, average absolute relative error and skill score). Researchers have also focused on regional performance analysis of GCMs in the Pacific Northwest USA. Werner (2011) evaluated 22 GCMs from CMIP3 datasets using various performance metrics generated by work from other groups namely Pincus et al. 2008; Pierce et al. 2009; Jost et al. 2012. Both global as well as regional performance analysis was used to have robust results. They used results of those studies and determined several decision factors. Some factors were based on statistical measures obtained from GCMs, and some considered availability and performance of GCMs in other studies. Recently, Rupp et al. (2013) used 41 CMIP5 GCMs and 24 CMIP3 GCMs and evaluated each model's simulation for the Pacific Northwest USA with observational gridded dataset. They defined 19 performance metrics and evaluated each model according to their performance on those metrics.

In the present study, we have analyzed the performance of 20 GCMs from CMIP5 dataset based on their performance in accordance with historical gridded observational data (Livneh et al. 2013) over Columbia River Basin in Northwest USA. We have based our analysis on precipitation and temperature, since precipitation is the main input for hydrological models, and temperature plays a key role in the estimation of evaporation and evapotranspiration (Woldemeskel et al. 2012). A wide range of statistical methods have been applied on the raw simulations from GCMs and gridded observational data to assess their performance based on the properties/attributes captured by the particular statistical method in the historical period of 1970-2000. Nevertheless, our evaluation method is general (based on different statistical properties of data i.e. univariate and multivariate analysis) and can be used in any other regions to evaluate climate models. The

motivation for this study included analysis of daily data, which is reported in results section, but we have also performed the analysis on monthly and seasonal (summer and winter) dataset. For brevity, only daily dataset statistics are reported in results section and same could further be used in hydrological analysis on daily time scale. Other temporal scales are reported only for the final evaluation matrix. Effect of change of observational dataset was also studied by evaluating the raw GCM simulations with Abatzoglou (2013) gridded observational dataset in the study area.

Results of this study were utilized in parallel efforts to assess the impacts of climate change and global warming on characteristics of climatic variables over Columbia River Basin (CRB). Rana and Moradkhani (2015) analyzed spatial, temporal, and frequency changes of future precipitation and temperature in CRB using this set of selected GCMs. The application of 40 different downscaled models/scenarios for various timescales has provided insight into probable changes in future climate. Demirel and Moradkhani (in press) applied Bayesian Model Averaging to reduce the uncertainty in GCM predictions for studying the seasonality and timing of historical precipitation over Columbia River Basin. Their results identified the changes in seasonality and persistence of extreme precipitation events for the study region.

The paper is divided into 6 sections, introduction followed by description of study area and data. This is followed by description of univariate and multivariate statistical procedures used for analysis and results, discussion and finally summary and conclusion is outlined in section 5 and 6.

Study Area and Data Used

Daily records of precipitation (P) and near surface temperature (T) in the study region (Figure 1) were collected for 20 GCMs (table 1) of the CMIP5 historical experiment (Taylor et al. 2012). The areal daily average for precipitation and temperature is calculated over the Columbia River Basin

(Figure 1) for each GCM along with other accumulated temporal scales of monthly and seasonal (summer and winter) datasets (accumulation from daily values). The GCM data is evaluated against gridded daily dataset acquired from University of Washington (Livneh et al. 2013) (hereafter referred to as gridded observational data), which has a spatial resolution of 1/16 Deg., and is available for the historical period of 1970-2000. This is the most widely used (and reliable) dataset in study area. Gridded observational dataset (Abatzoglou 2013) from University of Idaho with spatial resolution of 1/24 Deg. was also used to study the effect of observational dataset on selection/evaluation of GCMs. GCMs and gridded observation data each have different spatial resolution and hence, they cannot be compared on grid scale without statistical manipulations, like interpolation. Therefore, spatial average values of GCMs and observation are used in all the analysis. Also, each method is applied separately on Precipitation and Temperature.

Methods

The performance evaluation matrix deployed in this paper is based on the ability of particular GCM to reproduce the statistical properties/attributes of the gridded observational data, and no direct comparison of time series is done for simulations and observations. We have not based the evaluation of models on a particular matrix/method as opposed to what is suggested by others (Hawkins and Sutton 2011; Deser et al. 2012a; Deser et al. 2012b; Deser et al. 2014). Instead we have reported the evaluation on a number of metrics to provide a broader basis for assessment and decision making on various time scales based on user interest. This would also help to remove subjectivity connected with regional/local properties or previous knowledge of the area concerned. Although, choice of relevant climate variables/spatial/temporal resolutions and ranges etc. would still be subjected to user discretion and not target study area. Thus, the process is objective and based only on the statistical properties of GCM data and that of gridded observational data and the

user need not have any prior knowledge of the area in concern, which in turn adds to the advantage of its application in any area. The performance of different GCMs in a particular method can also be investigated. Furthermore, it is possible to compare the ability of different methods as they address various statistical properties.

Various performance metrics have been proposed by researchers. Some of these metrics focus on the mean climatological state, whereas others are related to temporal variability (e.g. seasonal variations, yearly and decadal changes). Since there is no standard methodology to evaluate climate models, we chose metrics, which are statistically credible, and are able to examine the statistical characteristics of models in accordance with gridded observational data. The metrics compare the distribution properties of models (mean, variance, correlation, among others) as well as the trends and relative changes. Various metrics applied focus on certain statistical properties of the dataset itself. An overall of 10 metrics are employed to compare the performance of each model (and each temporal scale) with the gridded observational data; this is the basis of multi-criteria analysis. Thus, the end user has 40 metrics (4 temporal scales*10 evaluation methods) for each of the climatic variable, i.e. precipitation and temperature; total of 20 metrics for ranking GCMs on particular temporal scale. The metrics can be classified under univariate and multivariate statistical measures of performance.

Univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. The metrics that are used for univariate statistical analysis in the study are:

1. Mean
2. Standard deviation

3. Coefficient of Variation (CV)
4. Relative Change (Variability)
5. Mann-Kendall Trend
6. Kolmogorov-Smirnov test (KS-test)

Multivariate statistics is the form of statistics encompassing the simultaneous observation and analysis of more than one outcome variable in the dataset. The following multivariate techniques were applied in the study:

7. Principal Component Analysis (PCA) or Empirical Orthogonal Function (EOF)
8. Singular Value Decomposition (SVD) or Maximum Covariance Analysis
9. Canonical Correlation Analysis (CCA)
10. Cluster Analysis

All the metrics are applied on spatially averaged GCMs and observational datasets. For multivariate metrics, the analysis is performed after standardizing datasets. A brief explanation of methods along with the statistical properties they address is provided in the following paragraphs. The methods are applied for both precipitation and temperature separately. More detailed information about multivariate methods can be found in Bretherton et al. (1992).

3.1 Univariate Statistics

Mean of dataset refers to the central tendency either of a probability distribution or of the random variable characterized by that distribution; and Standard deviation measures the amount of variation or dispersion of data from average/mean. Calculating them will reveal how data is distributed, and the range that most of the average values occur. The coefficient of variation is defined as the ratio of the standard deviation σ to the mean μ, i.e. normalized measure of dispersion

of a probability/frequency distribution. It removes the dependency of standard deviation on the mean, and investigates the variability in relation to mean of population. In this study, coefficient of variation is calculated for all temporal scales of each GCM and also for gridded observational data. For temperature, CV is calculated using data in Kelvin.

Relative change (RC), in quantitative science, evaluates the relative difference or variability of models while taking into account sizes of things being compared. Since there is large variations in daily values of precipitation and temperature, relative change is only calculated at the yearly scale. Thus for each GCM, absolute annual RC is calculated in the study period for both variables. Then, average absolute RC over the entire period is used for ranking, and the GCM which has a similar average absolute RC to observation receives a higher score. Relative change of temperature is calculated using data in Kelvin. Large changes infer little or no consistency in precipitation/temperature between different years. Relative change removes the dependency of standard deviation on mean (Rana et al., 2013).

Trends- Mann-Kendall Test: The rank-based Mann-Kendall test is a non-parametric test i.e. independent of the statistical distribution of the data. The Mann-Kendall trend test is based on the correlation between the ranks of a time series and their time order. For more information, readers are referred to Belle and Hughes (1984) and Govindarajulu (1992). Ranking is performed using the test statistics (z-value) at the given significance level (95% in this case). Using test statistics, one can easily understand if the trend is positive or negative. Furthermore, since all datasets have the same length and the confidence interval is constant, significant test statistics value can be easily found. The results are analyzed as follows:

(a) If the test statistics obtained for gridded observational data is positive, models with positive and closer statistics to observation will receive a score of 5. Values calculated for other models are divided into 4 groups based on their test statistics, and ranking is performed based on the proximity to observational statistics.

(b) If the test statistics obtained for gridded observational data is negative, models with negative and closer statistics to observation will receive a score of 5. Values calculated for other models are divided into 4 groups based on their test statistics, and ranking is performed based on the proximity to observational statistics.

Kolmogorov-Smirnov test (KS-test) is one of the most useful and general non-parametric methods for comparing two samples to decide whether the samples come from a population with a specific distribution. The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution (in the two-sample case) or that the sample is drawn from the reference distribution (in the one-sample case). It is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples. KS-test is distribution free test and is based on looking at the maximum vertical distance between the ECDF of the two distributions. More information about KS test can be found at Huth and Pokorn (2004). In this study, the two-sample KS test is applied over samples of each GCM and observation, and for each case test statistics are calculated and used for ranking. This is done for all temporal scales on daily, monthly, and seasonal.

3.2 Multivariate Statistics

Principal Component Analysis (PCA) or Empirical Orthogonal Function (EOF): It simplifies (using orthogonal transformation) the complex interrelationships in a dataset by constructing one

238 or few variables, which enable easier assessment of the relationships (Moradkhani and Meier 2010;
239 Rana et al. 2012). PCA maximizes the variance explained by weighted sum of elements in two or
240 more fields by recognizing linear transformations of the dataset that describes the variance as much
241 as possible in a few number of variables. PCA specifies the relationship among various modes of
242 variability by separating the modes in time series of different fields. It searches for basis vectors
243 that can describe the behavior of multiple variable metrics (Nishii et al. 2012). PCA isolates the
244 modes of variability observed in time series of different fields and gives their relationships in
245 separate modes. In this study, PCA is performed on standardized data of each GCM and the gridded
246 observational data for all temporal scales. Desired components are selected, and eigenvalues of
247 each model are compared to the eigenvalue of gridded observational data.

248 Singular Value Decomposition (SVD) or Maximum Covariance Analysis: SVD, a matrix
249 operation, is applied to asymmetric or not squared matrices in the diagonalization of PCA. It
250 provides the spatial patterns from the two fields that explains most of the covariance between them
251 and thus also called Maximum covariance analysis. Maximizing the covariance between linearly
252 related variables makes SVD neutralize the linear combination of variables, which seem to be
253 linearly related to each other. The principal difference in both the techniques applied here is
254 maximization of variance in PCA whereas we maximize covariance of predictor and predictand in
255 case of SVD. For more information and detailed explanation of SVD refer to Bretherton et al.
256 (1992). Covariance explained by predictor in the predictand field in a particular mode is used to
257 compare the relative significance of certain mode in the expansion. The correlation coefficient
258 between the predictor and predictand provides information about how strong the two fields are
259 related to each other (Wallace et al. 1992). In this study, SVD is applied to the cross-covariance
260 matrix of the standardized GCMs and observation, where gridded observational data is assumed

as the predictand, and the models are treated as predictors. Heterogeneous correlation map—defined as the correlation between model values and the first expansion coefficient, obtained from each model is taken into account, and the model with higher correlation gets a higher score, performed for each temporal steps. It should be noted that there is no direct comparison of the time series itself, but instead with attributes of the time series, expansion series, and weight vectors, on various temporal scales i.e. daily, monthly, and seasonal.

Canonical Correlation Analysis (CCA): CCA measures the linear relationship between two multi-dimensional variables i.e. of the cross-covariance matrices of the data. It finds two optimal bases (one for each variable) according to correlation, and finds the corresponding correlations. CCA tries to find the bases in which correlation matrix between the variables is diagonal and the correlations on the diagonal are maximized. It might be treated as a special form of empirical orthogonal function (EOF) analysis, where it can describe the correlation between predictor and predictand more comprehensively using various modes in it (Barnston and Ropelewski 1992). CCA is applied on standardized data in the present study. Since CCA is a linear technique, its applicability is narrowed to relations wherein predictand and predictor have the same response. Therefore, it brings information about small perturbations than to assess strong nonlinear relations (Wójcik 2014).

More information and detailed explanation of CCA can be found in Wilks (2011). The differences among PCA, SVD, and CCA can be found at Bretherton et al. (1992). Spatial canonical correlations obtained from CCA performed on each GCM and gridded observational data is used to rank them accordingly for each of the temporal scales in consideration. More details about ranking and criteria used can be found in section 3.3.

Cluster Analysis: It methodologically tries to separate objects in various groups each having more similarities together than with other clusters (Bratchell 1989). Cluster analysis is an appropriate method to classify climate zones, and is becoming more practical in atmospheric research studies (Unal et al. 2003). It graphically depicts the relation among various observations by producing dendograms. Dendograms (also called cluster trees) present a number of levels of (dis)similarities, and place observations in different levels according to their similarities. Here we have constructed clusters from the agglomerative (start with points as individual clusters and, at each step, merge the closest pair of clusters) hierarchical clustering as generated by the linkage function. We have used flexible linkage method to classify models since it seemed to work more reasonable with climatic and hydrological datasets, based on literature review. Each GCM has a linkage distance to connect to the gridded observational data. These distances are extracted for models and they are ranked accordingly for all temporal scales. More information about cluster analysis can be found in Wilks (2011).

A summary of the type/characteristic of datasets used to perform each method is provided in table 2.

3.3 Model ranking

Evaluating GCMs with various statistical tests helps investigate the advantages and caveats of each model/GCM from various statistical aspects in respect to gridded observational dataset. However, it brings some challenges to interpret the results. In some studies, researchers have eliminated those metrics, and provided their ranking with some of their previously chosen metrics (Werner, 2011). Whereas in some researches, weights have been assigned to each method and then final ranking is presented based on weighted methods (Rupp et al., 2013), with some working on previous

knowledge of the area to eliminate the method/model which in turn brings subjectivity in the scenario.

In this study, we have chosen metrics which evaluate the important aspects of climate data and are not significantly redundant. Although some of the methods might seem similar and evaluate same feature, they are targeting different aspects of datasets. Moreover, in each method, outlier GCMs are excluded with lowest rank assigned, and thus ranks obtained by each method is checked to avoid possible overrating of a model. In other words, considering one method, if one of the GCMs performs poorly, it is first excluded to provide a more meaningful comparison among the GCMs. This can be verified in the figures of final rankings in results section. Metrics are treated equally to treat the methods objectively, since adding weights will be based on another assumption, which may increase the uncertainty. We have provided results of each metric for all models for further use in certain applications and for all the temporal scale in consideration. Use of various temporal scales, i.e. daily, monthly, and seasonal provides a wide range of array for stakeholders and decision makers to make decision based on the utility. It also contributes to study of various low frequency events that are not prominent on daily scale but are part of monthly and seasonal scale data, thus accommodating all the possible ranges of variability explained by the data. Daily scale results are emphasized throughout the study due to importance of daily data in driving hydrological models and analysis. Rankings are based on assigning scores of 1 to 5 for each metric. In other words, performance of each model will be compared to the gridded observational data and consequently it will receive a score of 1 to 5 on each metric, where 1 shows the least efficiency and 5 represents the best performance on the metric. Overall ranking is the summation of scores obtained for precipitation and temperature in each method for a GCM. Average of overall ranks of

each GCM is calculated and will be used to select GCMs on each time scale. Representative results are explained in the result section for each of the methodology applied.

Results

Evaluating each metric and studying the performance of models in them is an important aspect of investigating the overall performance for both variables. Therefore, results for each metric will be evaluated and discussed in the following sections and eventually ranking would be done based on results of each metric. An overall ranking based on averaged score for both variables would be provided thereafter.

4.1 Raw GCM Simulations and Gridded Observational Data

Before evaluation of GCM simulations, it is vital to explain the data itself and its characteristics. Boxplots and violin plots are the tools used to investigate/illustrate the raw GCM simulations and gridded observational data. Figure 2 illustrates boxplots of precipitation and temperature. In the figure, plots A, B, and C are depicting GCM raw simulation for each model and the gridded observational data for daily and seasonal precipitation; plots D, E, and F are representing temperature for the same timescales. In the figure, outliers are specified using markers with red color along with median in center of box and 25^{th} and 75^{th} percentiles marking box boundaries. Since there are many days with no precipitation, the median is around zero, and therefore, most of the data is assumed as outliers. However, for temperature (Fig. 2c), median and quartiles are clearly obvious and models can easily be compared to the gridded observational data. From figure 2, for daily precipitation (plot A), it can be noted that most of the models are overestimating the precipitation values and underestimating the dry days, all the models have median, along with 25^{th} and 75^{th} percentiles, higher than the gridded observational dataset. Overestimating precipitation is

more noticeable in warm season (plot B) when all GCMs are predicting higher values, and only CanESM2 shows low bias. Precipitation prediction of GCMs has less bias in cold seasons (plot C). On the other hand, for daily temperature, median and quartiles seem to be well predicted by climate models (Fig.2 D), and the outliers are only towards the lower temperature ranges. The observation has narrow box and fewer outliers than the GCM simulations and median is always equal or lower than the GCM simulations (Fig.2D). This simply indicates that GCMs tend to predict more extreme cold temperatures than observation. For seasonal temperature (Fig.2 E and F), most GCMs seem to underestimate observed temperature of cold season, and they show less bias in warm season.

4.2 Mean, Standard Deviation, Coefficient of Variation and Relative change

Results of Mean, standard deviation, coefficient of variation and relative change for each GCM and gridded observational data are depicted in figure 3. The figure 3a and 3b represents the mean along with ±1 standard deviation of precipitation and temperature, respectively. It can be observed from the figure that the precipitation distribution of GCMs are strikingly different from that of the gridded observational dataset, usually overestimating the precipitation and underestimating the dry days which can be attributed to drizzle effect in climate models (Beven, 2011). Thus, mean of gridded observational dataset is lower than all the GCMs and accordingly the spread (standard deviation) of dataset. Proximity of mean and standard deviation of each of the GCM is compared to that of observational mean and standard deviation to rank the models (from 1-5) consequently. However, the temperature distribution is in line with gridded observational data with GCM depicting higher spread than the latter. Similarly, mean and standard deviation proximity of the GCM is evaluated against the gridded observational data for ranking. Fig. 3c, CV for precipitation and temperature are specified with blue and red markers, respectively. The far right values (number

21) present results of gridded observational data. For precipitation, CV for gridded observational dataset is always higher than GCMs whereas 4 models have higher CV than the latter in case of temperature. As mentioned in methodology section, for temperature, CV and RC are calculated using data in Kelvin, since we have non-zero (negative) values in the region. Consequently, models with close proximity to gridded observational dataset would be ranked higher. Relative change shows the inter-annual variations of each variable (Fig. 3d and 3e for precipitation and temperature, respectively). Thus, it might be positive for some years and negative in some other years. The RC for precipitation of gridded observational dataset shows higher spread in boxplot suggesting higher relative change during years than in GCMs whereas it is opposite for temperature wherein the gridded observations have lower relative change than GCMs. The absolute value RC is calculated for each year and then average absolute RC for each GCM and gridded observation is calculated for evaluation. Proximity of CV and RC values of GCMs to gridded observational dataset is used for ranking from 1-5.

4.3 Mann-Kendall test

Trend analysis is performed using Mann-Kendall for gridded observational data and for each model. Values for models are then compared to the value obtained for gridded observational data. Results of trend analysis of precipitation and temperature are tabulated in table 3. In the table, results from Mann-Kendall test on daily precipitation and temperature are shown in the first two columns, followed by decadal change in each variable (using annual data) presented in the last column. Daily results are used for ranking on daily timescale and decadal change is shown to provide more knowledge about the study area. Results from daily Mann-Kendall test on observation dataset show significant positive trend for both precipitation and temperature dataset. Thus for both variables, all the models showing positive trend would be ranked higher relative to

negative trend ones in the period under consideration. For precipitation, BCC_CSM1_1m, BNU_ESM, GFDL_ESM2G, GFDL-CSM5A-LR, GFDL-CSM5B-LR, and MIROC5 gets the highest ranking of 5 due to positive, significant trend and proximity to statistics of observational dataset and other models are ranked accordingly. Whereas, for temperature, only BCC_CSM1_1 and CanESM2 gets ranking 5 and other models are ranked consequently in comparison to observed statistics. In both cases, i.e. for precipitation and temperature, models with negative trends would receive a least score. It can be observed from the table that many of the models are showing significant trend at 99% as well (p values ≤ 0.01) for both the variables and only few models do not show any trend in the dataset.

4.4 Kolmogorov-Smirnov test (KS-test)

KS-test is performed for gridded observational data and each GCM simulations at all the temporal scales. KS-test statistics are then compared to provide model ranking, on all temporal scales, for both the variables i.e. precipitation and temperature. Results of KS-test are presented in table 4. Since all the simulations, for both precipitation and temperature, rejected null hypothesis i.e. no time series were same at desired alpha, we have considered statistics of the test to evaluate the models in comparison to observational gridded data. The p-values were significantly very small in all the cases to develop a rational comparison of observational data and simulations. Therefore, test statistics are extracted and used for rankings. The statistics close to zero are better representation of the observational dataset and result in lower p-value and thus ranked higher. As can be seen from table 4, for daily precipitation data, BCC_CSM1_1m, CCSM4, CSIRO_Mk3, HadGEM2-CC, HadGEM2-ES, IPSL-CM5A-MR, and NorESM1-M are closest in respect to maximum vertical distance of empirical distribution function to observational gridded data and thus given highest ranking and vice versa for MIROC-ESM and MIROC-ESM-CHEM, with

418 farthest from observational data for precipitation. Whereas, in case of temperature daily data,
419 GFDL-ESM2M, INMCM4, IPSL-CM5A-LR, IPSL-CM5B-LR, and MRI-CGCM3 are closest as
420 opposed to HadGEM2-ES and MIROC5, being the farthest ones to observational dataset. Same
421 procedure was applied to rank the models on other temporal scales of monthly and seasonal.

422 **4.5 Principal Component Analysis (PCA)**

423 The percentage of variance explained by each component in PCA is studied and presented as pareto
424 graph in figure 4c and 4d for precipitation and temperature, respectively. Different components
425 describe different features in each variable and can be used for various purposes. Depending on
426 variance explained (acceptable level of variance explained based on user interest) by each mode
427 of PCA, user can decide on the number of modes to be used in analysis of the data. In this study,
428 for precipitation the first component explained about 10% of total variance whereas for
429 temperature it was about 89% of the total variance. The local variance (squared correlation
430 between the GCM simulation and the gridded observational dataset) in first component of PCA is
431 used for ranking the models for both variables. Performance of all the models in accordance to
432 squared correlation with gridded observational data is classified in 5 equal intervals, resulting in a
433 score of 1-5 based on their performance. Results for precipitation and temperature are graphically
434 presented in figure 4a and 4b, respectively. Figure 4a represents the various models in relation to
435 averaged gridded observational data in various components of PCA for precipitation and 4b
436 represents the same for temperature. The relative length i.e. distance from center for a particular
437 component (component 1 in this case) of the GCMs defines the relative proximity with the gridded
438 observational data. When the GCM is closer to gridded observational dataset (e.g.
439 BCC_CSM1_1m, CCSM4, and INMCM4 for precipitation), they will receive higher ranking, as
440 compared to ones which are distant from the same (e.g. IPSL-CM5B-LR, MIROC5, MIROC-

441 ESM, and MIROC-ESM-CHEM). Similarly, for temperature, GCMs (BCC_CSM1_1m,
442 BNU_ESM, CANESM2, CCSM4, GFDL_ESM2G, GFDL_ESM2M, HadGEM2-CC, INMCM4,
443 IPSL-CM5A-LR, IPSL-CM5A-MR, IPSL-CM5B-LR, MRI-CGCM3, and NorESM1-M) closer to
444 gridded observational dataset receives higher ranks and vice-versa (Fig. 4b).Same procedure is
445 performed for other temporal scales of monthly and seasonal.

446 **4.6 SVD and CCA**

447 Heterogeneous correlation representing maximized covariance between the predictand and
448 predictor is calculated for each GCM using SVD and is used to rank models (table 5). GCMs with
449 higher heterogeneous correlation represents similar properties/attributes with reference to gridded
450 observational data and are more suited for the study area. From table 5, it can be inferred that
451 BCC_CSM1_1, BCC_CSM1_1m, and BNU_ESM presents highest heterogeneous correlation and
452 thus receive the highest ranking for precipitation dataset, with IPSL-CM5A-LR, MIROC-ESM,
453 MIROC-ESM-CHEM, and NorESM1-M receiving the lowest. For temperature, CNRM_CM5,
454 IPSL-CM5A-MR, andMIROC5 are in close proximity to gridded observational dataset (based on
455 heterogeneous correlation) and are ranked highest; whereas HadGEM2-ES, MIROC-ESM, and
456 MIROC-ESM-CHEM are on the other end of ranking.

457 Similarly, CCA results were analyzed based on the similar property of GCMs and gridded
458 observational dataset. In other words, after calculating anomalies of a matrix (GCMs) versus
459 gridded observational data, and calculating PCA of the predictand, predictor canonical spatial
460 function is computed. Values of canonical spatial function (SF) are used to rank models. Models
461 with higher SF values will receive a higher score (table 6). The range of SF across the models is
462 divided in 5 groups and the highest value group will receive a score of 5. For CCA, BCC_CSM1_1,
463 CCSM4, HadGEM2-CC, INMCM4, IPSL-CM5A-MR, and MIROC5 receives the highest ranks

for precipitation dataset, whereas MIROC-ESM and MIROC-ESM-CHEM are ranked lowest. For temperature, BCC_CSM1_1, CanESM2, HadGEM2-CC, IPSL-CM5A-LR, and MRI-CGCM3 are amongst the higher ranked ones, whereas INMCM4, IPSL-CM5B-LR, and MIROC-ESM-CHEM receives the lower ranks.

4.7 Cluster Analysis

Results for cluster analysis are presented in figure 5a and 5b for precipitation and temperature, respectively. The plots/dendograms are showing different clusters among models and gridded observational data. Dendograms represents both the cluster-subcluster relationships and the order in which clusters are merged or split. Cluster group and linkage distance are important in determining the relative likelihood of models to represent the gridded observational dataset. As it can be interpreted from the dendograms, models have been distributed in several clusters which in turn are connected to each other in the last merged row. In figure 5, the plots show the value of linkage distance in accordance to the merged cluster indices, which are linked in pairs to form binary tree. Linkage distance reflects the degree of difference between branches i.e. longer lines indicate greater difference, principle applied to rank the models. Models which are in the same cluster with the observation (close proximity), are better performing than others. Similarly, the lesser the linkage distance of the model to observation, the higher the performance of the model, and the model receives a better rank. For precipitation (figure 5a), it can be observed that BCC_CSM1_1m (number 2) is in the closest proximity and belongs to same cluster as gridded observational dataset followed by BCC_CSM1_1 (number 1) and CCSM4 (number 5), forming the next closest cluster, and thus would receive highest rankings. The scale of model ranks are classified into 5 classes and ranked on the basis of same. IPSL-CM5B-LR, MIROC-ESM, and MIROC_ESM_CHEM are farthest forming a farthest cluster based on linkage distance and thus

receives lowest scores for precipitation dataset. For temperature (figure 5b), GFDL_ESM2G, IPSL-CM5A-LR, CCSM4, and IPSL-CM5A-MR are in close proximity to gridded observational dataset (ranked highest), whereas MIROC-ESM, MIROC-ESM-CHEM, BNU_ESM, BCC_CSM1_1m, and MRI-CGCM3 forms the farther clusters and thus ranked lower.

4.8 Overall Performance

Models performances were assessed in 10 metrics for precipitation and temperature, totaling to 20 metrics for each of the temporal scales in consideration, and each model received a score of 1-5 in each metric. Overall ranking, summation of ranks for precipitation and temperature, is provided using all 20 metrics values for each of the temporal scales in consideration. Performance of models on all temporal scales and each metric is depicted in figure 6. From figure 6 and table 7 it can be inferred, based on average overall performance for daily temporal scale, that CCSM4, IPSL-CM5A-MR, INMCM4, IPSL-CM5A-LR, CanESM2, GFDL_ESM2G, BCC_CSM1_1, GFDL_ESM2M, IPSL-CM5B-LR, and MIROC5 are 10 best representative GCMs of the gridded observational dataset (in order of decreasing ranking) in the desired period for the study region. Similar rankings for 10 best representative models for monthly, and seasonal and dataset is provided in figure 6 and table 7. End users can choose to have their own set of models based on utility and time scale in consideration. It can be observed from table 7 that GFDL_ESM2G, CCSM4, IPSL-CM5A-MR, and CanESM2 are among those selected at daily, monthly, summer, and winter temporal scales.

Discussion

The changing climate requires an investigation on understanding of its effects and causes on the environment and hydrological cycle. One of the most used resources for this purpose now-a-days,

509 are GCMs which represent the conditions of climate over the globe with predictions for future
510 scenarios. Each of these models has uncertainty associated in their predictions, and as they are
511 large-scale (coarse resolution), they might have different performances in regional scales/finer
512 resolution. Thus, there is a demand to investigate the performance of global models on regional
513 scales. We also intended to study the effects of observation dataset on the GCM selection
514 procedure thus we changed the gridded observational dataset with another gridded observational
515 dataset (Abatzoglou 2013). Similar statistical evaluation and ranking was performed for raw GCM
516 and the changed observational dataset on daily, monthly, and seasonal temporal scale, results are
517 presented in figure 7. It can be noted that the 10 best representative models (with changed gridded
518 observation) includes BCC_CSM1_1, GFDL_ESM2M, CCSM4, GFDL_ESM2G, MIROC5,
519 CanESM2, IPSL-CM5A-MR, IPSL-CM5B-LR, IPSL-CM5A-LR, and MIROC-ESM (in order of
520 decreasing ranking). On comparison, at daily temporal scale, with raw simulation evaluation based
521 on Livneh et al. (2013) gridded observational dataset to that of Abatzoglou (2013) it was found
522 that 9 of the models are represented in both the procedures, with only INMCM4 excluded in later
523 one (which is ranked 11^{th} in changed observational evaluation). It can be concluded that the
524 observational dataset have minimal effect on selection of GCMs which could be attributed to
525 averaging of climatic variables in the study area, making the two observational dataset comparable
526 to each other. Thus it becomes increasingly important to select the observational dataset based on
527 the physical representation in the study area for such analysis.

528 Various statistical methods, temporal scales and dataset have been used to analyze the range of
529 selection possibilities of GCMs in the study area. The advantages of this approach, among others,
530 include easy classification of models, quantitative-based and objective ranking. Hence, less
531 subjectivity is included in the results and users are not expected to be familiar with the study area

in question and thus could be applied in any study area. Moreover, the proposed methods are easy to perform and are handy in understanding the distribution and various statistical properties of the data. It is also suggested in literature (Werner, 2011) to remove multiple models from the same climate institutions so as to deal with the uncertainty associated with them, but that would not suffice the goal of the study in authors' opinion, as we are evaluating the model and not the institution for the capability of prediction. Moreover the GCMs from same institutions have different model setup and thus different simulations from each of them. Also, the results of study have indicated that models from same institution have behaved differently towards the analysis performed in the study.

It is also worth exploring the spatial aspects of the GCM selection in comparison to observational dataset. As pointed out in table 1, most of the GCMs have very varied spatial resolution and thus we adopted spatial average approach to evaluate GCMs against observational gridded data in Columbia River Basin. Depending on the scale and intent of the study same procedures can be applied on finer resolution of spatial data, as per availability of fine resolution observations, to compare the two sets. It would be interesting to study the spatial aspects based on elevation and various hydrological regimes in the study area, depending on the intent of the study.

CMIP3/CMIP5 simulations have long been used in various studies to evaluate different characteristics of climate change on humans and environment. Characteristics/trends of extreme events have been assessed in various studies (Mallakpour and Villarini, 2015; Najafi and Moazzami, 2015), some of which have used GCM data. The methodology proposed in this study can be applied on daily to multi-day extreme precipitation and temperature data to evaluate GCMs according to their performance in regard to extremes of these variables. Selecting appropriate GCMs would reduce uncertainty of future predictions (in comparison to other GCMs in the study

555 region), which is crucial for studying extreme conditions (e.g. floods or droughts), when the least
556 uncertainty is desired. GCM predictions have been used in various studies to detect and attribute
557 hydroclimate changes to human effects (Najafi et al, 2015; Zhang et al. 2013). Eventually,
558 selection of GCMs based on statistical attributes to evaluate various impacts according to the study
559 purpose would help reduce the various uncertainties associated with the larger GCM scale and be
560 helpful in large scale planning and management.

561 Daily dataset from CMIP5 has helped in a more robust analysis, and compare models with more
562 reliability. Metrics used in the present study are among the common statistical methods used in
563 several previous researches, and proved to work fine. Utilizing a variety of methods, each focusing
564 on a certain aspect of performance, along with using two most common climatic parameters brings
565 robustness to the analysis. Different temporal scales are considered in the study for various
566 stakeholder interest and user based analysis. Evaluation of the results of the present study reveals
567 that models generally perform better in temperature than in precipitation and a variable. This is
568 mainly because of the more stable nature of temperature which makes it easier to predict. Models
569 seemed to work differently in various methods. This might infer that models do not have high
570 correlation with each other. Finally, overall scores obtained by GCMs can be used for model
571 averaging or multi-modeling e.g.,(Najafi et al. 2011; Madadgar and Moradkhani 2014). In other
572 words, overall score of GCMs can be standardized and used as the weights applied in weighted
573 averaging. However, since this study is done using spatially averaged data and multi-modeling is
574 usually done at grid scale, it is not suggested to use the scores gathered here for weighted
575 averaging. Instead, one can first downscale all GCMs to a fixed spatial resolution and then apply
576 the methods proposed in this study and use results of each grid to calculate weights for GCMs
577 (Najafi and Moradkhani 2015).

Summary and Conclusion

Historical data for 20 GCMs from CMIP5, as well as gridded observational data were acquired and accumulated for different temporal scales of daily, monthly and seasonal (summer and winter). GCMs were evaluated with respect to their performance in simulating the climate in Columbia River Basin for historical period. Generally, all GCMs work fairly well in simulating temperature. However for precipitation, GCMs had various behaviors. This is mainly because the average rate of daily variations in precipitation is higher than temperature (e.g. considering two consequent days, one with no precipitation, the other one with heavy rain).

Utilizing daily data for 30 years, 10 metrics for 2 different parameters and different temporal scales have helped in robust assessment of models. Several metrics were chosen to investigate various aspects of model statistical properties. All metrics were treated equally and no weights were applied to the results of each metric to decrease the uncertainties. In general, GCMs usually behave differently in various methods, and no fixed methodology is presented to evaluate them. It is up to the research and the purpose of study to conduct a methodology and assess GCMs. The GCMs were also evaluated against different set of gridded observational dataset to study the effect of same on selection procedure. The presented methods can be applied/used for bias correction of the raw GCM data along with any or the statistical and dynamic downscaling method before using them in any study. This would help reduce the uncertainty in the model data. The present research should be considered as qualitative and that could be employed in dealing with GCMs data which in turn is driven by statistical properties of the data itself, which are often used in the field.

Acknowledgement

599 Partial financial support for this study was provided by the DOE-BPA (cooperative agreement
600 00063182). The authors would also like to acknowledge the World Climate Research Programme's
601 Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate
602 modeling groups (listed in table 1 of this paper) for producing and making available their model
603 outputs. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and
604 Intercomparison provides coordinating support and leads development of software infrastructure
605 in partnership with the Global Organization for Earth System Science Portals.

606 **References**

607 Abatzoglou JT (2013) Development of gridded surface meteorological data for ecological
608 applications and modelling. Int J Climatol 33:121–131. doi: 10.1002/joc.3413

609 Barnston a. G, Ropelewski CF (1992) Prediction of ENSO episodes using canonical correlation
610 analysis. J. Clim. 5:1316–1345.

611 Belle G, Hughes JP (1984) Nonparametric Tests for Trend in Water Quality. Water Resour Res
612 20:127–136. doi: 10.1029/WR020i001p00127

613 Beven, K. J. (2011). *Rainfall-runoff modelling: the primer*. John Wiley & Sons.

614 Bratchell N (1989) Cluster Analysis. 6:105–125.

615 Bretherton CS, Smith C, Wallace JM (1992) An Intercomparison of Methods for Finding Coupled
616 Patterns in Climate Data. J. Clim. 5:541–560.

617 Buser CM, Künsch HR, Lüthi D, et al (2009) Bayesian multi-model projection of climate: Bias
618 assumptions and interannual variability. Clim Dyn 33:849–868. doi: 10.1007/s00382-009-
619 0588-6

620 Chiew FHS, Teng J, Vaze J, Kirono DGC (2009) Influence of global climate model selection on
621 runoff impact assessment. J Hydrol 379:172–180. doi: 10.1016/j.jhydrol.2009.10.004

622 Christensen JH, Kjellström E, Giorgi F, et al (2010) Weight assignment in regional climate models.
623 Clim Res 44:179–194. doi: 10.3354/cr00916

624 Demirel, M., and H. Moradkhani (in press) Assessing the Impact of CMIP5 Climate Multi-
625 Modeling on Estimating the Precipitation Seasonality and Timing, Climatic Change.

626 Deser C, Knutti R, Solomon S, Phillips AS (2012a) Communication of the role of natural
627 variability in future North American climate. Nat Clim Chang 2:775–779. doi:
628 10.1038/nclimate1562

629 Deser C, Phillips A, Bourdette V, Teng H (2012b) Uncertainty in climate change projections: The
630 role of internal variability. Clim Dyn 38:527–546. doi: 10.1007/s00382-010-0977-x

631 Deser C, Phillips AS, Alexander M a., Smoliak B V. (2014) Projecting North American climate
632 over the next 50 years: Uncertainty due to internal variability. J Clim 27:2271–2296. doi:
633 10.1175/JCLI-D-13-00451.1

634 Fowler HJ, Blenkinsop S, Tebaldi C (2007) Review: Linking climate change modelling to impacts
635 studies: recent advances in downscaling techniques for hydrological modelling. Int J Climatol
636 27:1547–1578. doi: 10.1002/joc

637 Govindarajulu Z (1992) Rank Correlation Methods (5th ed.). Technometrics 34:108. doi:
638 10.1080/00401706.1992.10485252

639 Hawkins E, Sutton R (2011) The potential to narrow uncertainty in projections of regional
640 precipitation change. Clim Dyn 37:407–418. doi: 10.1007/s00382-010-0810-6

641 Hintze JL, Nelson RD (1998) Violin plots: A box plot-density trace synergism. Am. Stat. 52:181–
642 184.

643 Huth R, Pokorn L (2004) Parametric versus non-parametric estimates of climatic trends. Theor
644 Appl Climatol 77:107–112. doi: 10.1007/s00704-003-0026-3

645 Johnson F, Westra S, Sharma A, Pitman AJ (2011) An assessment of GCM skill in simulating
646 persistence across multiple time scales. J Clim 24:3609–3623. doi: 10.1175/2011JCLI3732.1

647 Jost G, Moore RD, Menounos B, Wheate R (2012) Quantifying the contribution of glacier runoff
648 to streamflow in the upper Columbia River Basin, Canada. Hydrol Earth Syst Sci 16:849–
649 860. doi: 10.5194/hess-16-849-2012

650 Livneh B, Rosenberg EA, Lin C, et al (2013) A Long-Term Hydrologically Based Dataset of Land
651 Surface Fluxes and States for the Conterminous United States: Update and Extensions*. J
652 Clim 26:9384–9392. doi: 10.1175/JCLI-D-12-00508.1

653 Madadgar S, Moradkhani H (2014) Improved Bayesian multimodeling: Integration of copulas and
654 Bayesian model averaging. Water Resour Res 50:9586–9603. doi:
655 10.1002/2014WR015965.Received

656 Mallakpour, I., Villarini, G. (2015). The changing nature of flooding across the central United
657 States. *Nature Climate Change, 5*(3), 250-254.

658 Maxino CC, McAvaney BJ, Pitman AJ, Perkins SE (2008) Ranking the AR4 climate models over
659 the Murray-Darling Basin using simulated maximum temperature, minimum temperature and
660 precipitation. Int J Climatol 28:1097–1112. doi: 10.1002/joc

661 Miao C, Duan Q, Yang L, Borthwick AGL (2012) On the Applicability of Temperature and
662 Precipitation Data from CMIP3 for China. PLoS One 7:1–10. doi:
663 10.1371/journal.pone.0044659

664 Moradkhani H, Baird RG, Wherry S a. (2010) Assessment of climate change impact on floodplain
665 and hydrologic ecotones. J Hydrol 395:264–278. doi: 10.1016/j.jhydrol.2010.10.038

666 Moradkhani H, Meier M (2010) Long-Lead Water Supply Forecast Using Large-Scale Climate
667 Predictors and Independent Component Analysis. J Hydrol Eng 15:744–762. doi:
668 10.1061/(ASCE)HE.1943-5584.0000246

669 Najafi, M. R., Moazami, S. (2015). Trends in total precipitation and magnitude–frequency of
670 extreme precipitation in Iran, 1969–2009. *International Journal of Climatology*.

671 Najafi MR, Moradkhani H (2015) Multi-model ensemble analysis of runoff extremes for climate
672 change impact assessments. J Hydrol 525:352–361. doi: 10.1016/j.jhydrol.2015.03.045

673 Najafi MR, Moradkhani H, Jung IW (2011) Assessing the uncertainties of hydrologic model
674 selection in climate change impact studies. Hydrol Process 25:2814–2826. doi:
675 10.1002/hyp.8043

676 Najafi, M. R., Zwiers, F. W., & Gillett, N. P. (2015). Attribution of Arctic temperature change to
677 greenhouse-gas and aerosol influences. *Nature Climate Change*.

678 Nishii K, Miyasaka T, Nakamura H, et al (2012) Relationship of the Reproducibility of Multiple
679 Variables among Global Climate Models. J Meteorol Soc Japan 90A:87–100. doi:
680 10.2151/jmsj.2012-A04

681 Önol B, Bozkurt D, Turuncoglu UU, et al (2014) Evaluation of the twenty-first century RCM
682 simulations driven by multiple GCMs over the Eastern Mediterranean-Black Sea region. Clim
683 Dyn 42:1949–1965. doi: 10.1007/s00382-013-1966-7

684 Pierce DW, Barnett TP, Santer BD, Gleckler PJ (2009) Selecting global climate models for
685 regional climate change studies. Proc Natl Acad Sci U S A 106:8441–8446. doi:
686 10.1073/pnas.0900094106

687 Pincus R, Batstone CP, Patrick Hofmann RJ, et al (2008) Evaluating the present-day simulation of
688 clouds, precipitation, and radiation in climate models. J Geophys Res Atmos 113:1–10. doi:
689 10.1029/2007JD009334

690 Raju K, Nagesh Kumar D (2014) Ranking of global climate models for India using multicriterion
691 analysis. Clim Res 60:103–117. doi: 10.3354/cr01222

692 Rana A, Uvo CB, Bengtsson L, Sarthi PP (2012) Trend analysis for rainfall in Delhi and Mumbai,
693 India. Clim Dyn 38:45–56. doi: 10.1007/s00382-011-1083-4

694 Rana A, Madan S, Bengtsson L (2013) Performance Evaluation of Regional Climate Models
695 (RCMs) in determining precipitation characteristics for Göteborg, Sweden. Hydrology
696 Research. doi:10.2166/nh.2013.160

697 Rana, A., and H. Moradkhani (2015) Spatial, temporal and frequency based climate change
698 assessment in Columbia River Basin using multi downscaled-Scenarios, Climate Dynamics,
699 DOI: 10.1007/s00382-015-2857-x.

700 Rupp DE, Abatzoglou JT, Hegewisch KC, Mote PW (2013) Evaluation of CMIP5 20th century
701 climate simulations for the Pacific Northwest USA. J Geophys Res Atmos 118:10884–10906.
702 doi: 10.1002/jgrd.50843

703 Samadi S, Wilson C a ME, Moradkhani H (2013) Uncertainty analysis of statistical downscaling
704 models using Hadley Centre Coupled Model. Theor Appl Climatol 114:673–690. doi:
705 10.1007/s00704-013-0844-x

706 Taylor KE, Stouffer RJ, Meehl G a. (2012) An overview of CMIP5 and the experiment design.
707 Bull Am Meteorol Soc 93:485–498. doi: 10.1175/BAMS-D-11-00094.1

708 Unal Y, Kindap T, Karaca M (2003) Redefining the climate zones of Turkey using cluster analysis.
709 Int J Climatol 23:1045–1055. doi: 10.1002/joc.910

710 Wallace JM, Smith C, Bretherton CS (1992) Singular Value Decomposition of Wintertime Sea
711 Surface Temperature and 500-mb Height Anomalies. J. Clim. 5:561–576.

712 Wang D, Hagen SC, Alizad K (2013) Climate change impact and uncertainty analysis of extreme
713 rainfall events in the Apalachicola River basin, Florida. J Hydrol 480:125–135. doi:
714 10.1016/j.jhydrol.2012.12.015

715 Werner AT (2011) BCSD Downscaled Transient Climate Projections for Eight Select GCMs over
716 British Columbia, Canada. Pacific Climate Impacts Consortium. University of Victoria.
717 Victoria. BC. 63 pp.

718 Wilks DS (2011) Statistical methods in the atmospheric sciences (Vol. 100). Academic press.

719 Wójcik R (2014) Reliability of CMIP5 GCM simulations in reproducing atmospheric circulation
720 over europe and the north atlantic: A statistical downscaling perspective. Int J Climatol
721 732:714–732. doi: 10.1002/joc.4015

722 Woldemeskel FM, Sharma a., Sivakumar B, Mehrotra R (2012) An error estimation method for
723 precipitation and temperature projections for future climates. J Geophys Res Atmos 117:1–
724 13. doi: 10.1029/2012JD018062

725 Zhang, X., Wan, H., Zwiers, F. W., Hegerl, G. C., & Min, S. K. (2013). Attributing intensification
726 of precipitation extremes to human influence. Geophysical Research Letters, 40(19), 5252-
727 5257.

728

729 **Tables**

730 Table 1.Models used in this study and their characteristics

S.No.	Model	Center	Atm. Resolution (Lon x Lat)	Vertical levels in Atm.
1	bcc-csm1-1	Beijing Climate Center, China Meteorological Administration	2.8 × 2.8	26
2	bcc-csm1-1-m	Beijing Climate Center, China Meteorological Administration	1.12 × 1.12	26
3	BNU-ESM	College of Global Change and Earth System Science, Beijing Normal University, China	2.8 × 2.8	26
4	CanESM2	Canadian Centre for Climate Modeling and Analysis	2.8 × 2.8	35
5	CCSM4	National Center of Atmospheric Research, USA	1.25 × 0.94	26
6	CNRM-CM5	National Centre of Meteorological Research, France	1.4 × 1.4	31
7	CSIRO-Mk3-6-0	Commonwealth Scientific and Industrial Research Organization/ Queensland Climate Change Centre of Excellence, Australia	1.8 × 1.8	18
8	GFDL-ESM2G	NOAA Geophysical Fluid Dynamics Laboratory, USA	2.5 × 2.0	48
9	GFDL-ESM2M	NOAA Geophysical Fluid Dynamics Laboratory, USA	2.5 × 2.0	48
10	HadGEM2-CC	Met Office Hadley Center, UK	1.88 × 1.25	60
11	HadGEM2-ES	Met Office Hadley Center, UK	1.88 × 1.25	38
12	INMCM4	Institute for Numerical Mathematics, Russia	2.0 × 1.5	21
13	IPSL-CM5A-LR	Institut Pierre Simon Laplace, France	3.75 × 1.8	39
14	IPSL-CM5A-MR	Institut Pierre Simon Laplace, France	2.5 × 1.25	39
15	IPSL-CM5B-LR	Institut Pierre Simon Laplace, France	3.75 × 1.8	39
16	MIROC5	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	1.4 × 1.4	40
17	MIROC-ESM	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	2.8 × 2.8	80
18	MIROC-ESM-CHEM	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	2.8 × 2.8	80
19	MRI-CGCM3	Meteorological Research Institute, Japan	1.1 × 1.1	48
20	NorESM1-M	Norwegian Climate Center, Norway	2.5 × 1.9	26

731

732

733

734

735

736

737 Table 2. Summary of data types/characteristics used in each method

Metric	Precipitation	Temperature
Mean	SA*	SA
Std dev	SA	SA
CV	SA	SA, Data in Kelvin
RC	SA, Annual timescale	SA, Data in Kelvin, Annual timescale
Mann-Kendall	SA	SA
KS-test	SA	SA
PCA	SA, Stdz**	SA, Stdz
SVD	SA, Stdz	SA, Stdz
CCA	SA, Stdz	SA, Stdz
Cluster	SA, Stdz	SA, Stdz
* SA: Spatially averaged over the study area ** Stdz: Standardized data		

738
739 Table 3. Mann-Kendall test statistics of both precipitation and temperature data (Values in bold are
740 significant at 95%). The last two columns indicate 30-year mean change of annual precipitation and
741 temperature for each model.

S. No.	Model	Precipitation		Temperature		30-year mean change of annual datasets	
		Z-Value	P-Value	Z-Value	P-Value	Prec. (%)	Temp. (°C)
1	BCC_CSM1_1	**-2.169**	**0.030**	**3.318**	**0.001**	-6.44	1.07
2	BCC_CSM1_1m	**3.589**	**0.000**	1.176	0.239	12.29	0.14
3	BNU_ESM	**2.057**	**0.040**	**2.506**	**0.012**	5.40	0.70
4	CanESM2	-1.651	0.099	**3.916**	**0.000**	-5.44	1.18
5	CCSM4	0.422	0.673	**5.009**	**0.000**	4.07	1.59
6	CNRM_CM5	0.658	0.510	**4.151**	**0.000**	0.70	0.86
7	CSIRO_MK3	1.910	0.056	1.073	0.283	7.46	0.14
8	GFDL_ESM2G	**2.232**	**0.026**	**2.381**	**0.017**	6.12	0.57
9	GFDL_ESM2M	**-2.151**	**0.032**	1.254	0.210	-4.16	0.08
10	HadGEM2-CC	1.281	0.200	0.372	0.710	3.66	0.07
11	HadGEM2-ES	-0.598	0.550	0.913	0.361	-0.75	0.37
12	INMCM4	-0.034	0.973	**4.268**	**0.000**	0.42	0.98
13	IPSL-CM5A-LR	**3.484**	**0.000**	1.558	0.119	9.71	0.32
14	IPSL-CM5A-MR	**-2.229**	**0.026**	**2.448**	**0.014**	-2.51	0.67
15	IPSL-CM5B-LR	**4.605**	**0.000**	-0.058	0.954	8.36	0.34
16	MIROC5	**2.325**	**0.020**	**4.470**	**0.000**	6.32	1.27
17	MIROC-ESM	-1.051	0.293	**6.819**	**0.000**	-1.24	1.81
18	MIROC-ESM-CHEM	0.775	0.438	**2.466**	**0.014**	1.28	0.30
19	MRI-CGCM3	0.923	0.356	0.383	0.702	1.91	0.07
20	NorESM1-M	**-2.629**	**0.009**	0.980	0.327	-4.67	-0.09

21	Gridded observational Data	**3.039**	**0.002**	**3.431**	**0.001**	6.95	0.61

Table 4. Statistics calculated in the two-sample Kolmogorov-Smirnov test. 95% confidence interval and unequal tail condition are taken as the assumptions in all cases. Smaller statistics value represent less difference in cumulative density function of model and observation, and thus is of more interest.

S. No.	Model	Precipitation	Temperature
1	BCC-CSM1-1	0.220	0.186
2	BCC-CSM1-1m	0.130	0.189
3	BNU-ESM	0.291	0.180
4	CanESM2	0.135	0.251
5	CCSM4	0.170	0.223
6	CNRM-CM5	0.262	0.160
7	CSIRO-Mk3	0.168	0.176
8	GFDL-ESM2G	0.230	0.212
9	GFDL-ESM2M	0.289	0.145
10	HadGEM2-CC	0.142	0.238
11	HadGEM2-ES	0.137	0.245
12	INMCM4	0.377	0.121
13	IPSL-CM5A-LR	0.236	0.144
14	IPSL-CM5A-MR	0.176	0.178
15	IPSL-CM5B-LR	0.179	0.143
16	MIROC5	0.224	0.272
17	MIROC-ESM	0.373	0.236
18	MIROC-ESM-CHEM	0.371	0.239
19	MRI-CGCM3	0.307	0.100
20	NorESM1-M	0.179	0.185

Table 5. Heterogeneous correlation calculated for each GCM by SVD

S. No.	Model	SVD	
		Precipitation	Temperature
1	BCC-CSM1-1	0.090	0.824
2	BCC-CSM1-1m	0.127	0.800
3	BNU-ESM	0.107	0.823
4	CanESM2	0.057	0.843
5	CCSM4	0.100	0.830
6	CNRM-CM5	0.071	0.862
7	CSIRO-Mk3	0.087	0.859
8	GFDL-ESM2G	0.067	0.845
9	GFDL-ESM2M	0.051	0.836
10	HadGEM2-CC	0.076	0.846
11	HadGEM2-ES	0.064	0.775
12	INMCM4	0.071	0.822
13	IPSL-CM5A-LR	0.024	0.833
14	IPSL-CM5A-MR	0.079	0.842
15	IPSL-CM5B-LR	0.058	0.820
16	MIROC5	0.053	0.858
17	MIROC-ESM	-0.025	0.687
18	MIROC-ESM-CHEM	-0.016	0.683
19	MRI-CGCM3	0.053	0.816
20	NorESM1-M	0.038	0.816

769 Table 6. Canonical spatial function (SF) calculated by CCA for each GCM

S. No.	Model	CCA	
		Precipitation	Temperature
1	BCC-CSM1-1	0.248	0.431
2	BCC-CSM1-1m	0.183	0.150
3	BNU-ESM	0.170	0.052
4	CanESM2	0.129	0.099
5	CCSM4	0.229	0.043
6	CNRM-CM5	0.146	0.043
7	CSIRO-Mk3	0.130	0.165
8	GFDL-ESM2G	0.113	0.029
9	GFDL-ESM2M	0.146	0.042
10	HadGEM2-CC	0.260	0.125
11	HadGEM2-ES	0.136	0.077
12	INMCM4	0.221	-0.180
13	IPSL-CM5A-LR	-0.024	0.101
14	IPSL-CM5A-MR	0.293	0.038
15	IPSL-CM5B-LR	0.024	-0.179
16	MIROC5	0.212	0.040
17	MIROC-ESM	-0.082	-0.105
18	MIROC-ESM-CHEM	-0.088	-0.151
19	MRI-CGCM3	0.172	0.107
20	NorESM1-M	-0.005	0.073

770

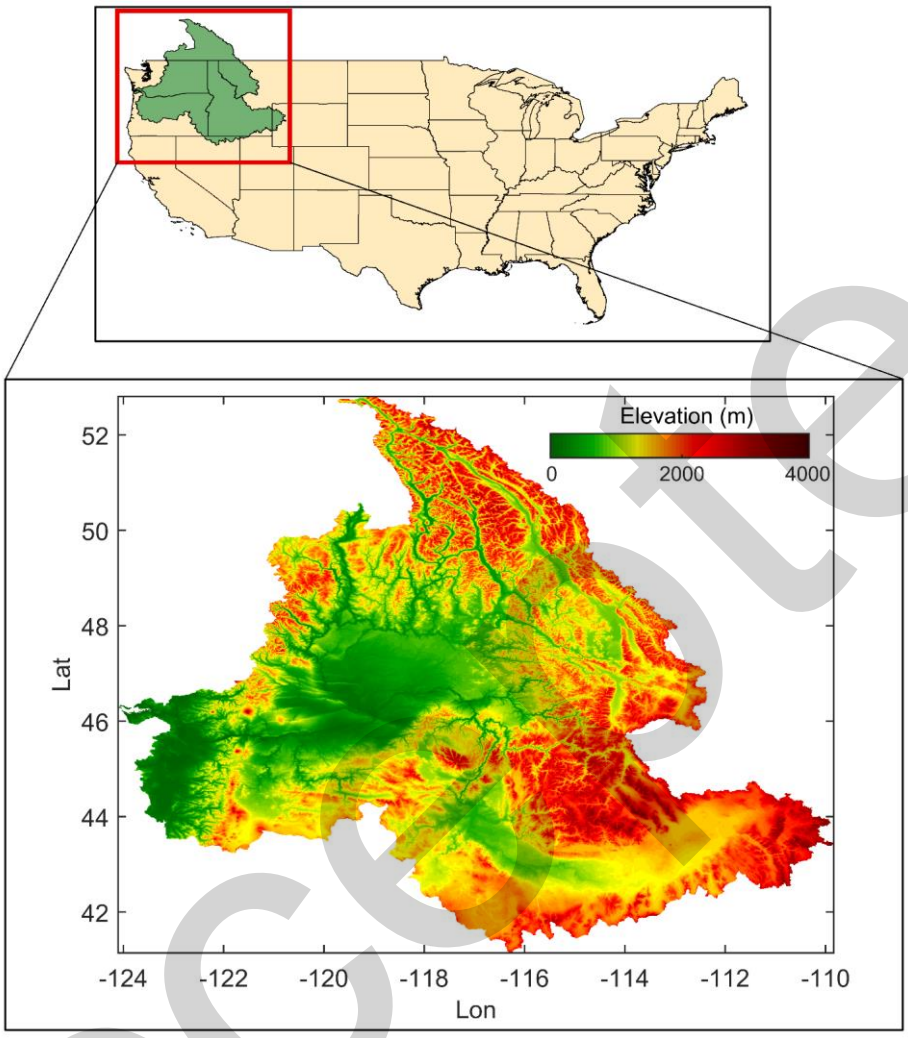
771 Table 7. List of top 10 models from 20 GCMs in the study for various temporal scales in order of decreasing
772 ranking. (Models in bold are common to all temporal scales)

No.	Daily	Monthly	Seasonal- Summers	Seasonal- Winters
1	**CCSM4**	**IPSL-CM5A-MR**	BCC_CSM1_1	INMCM4
2	**IPSL-CM5A-MR**	BCC_CSM1_1m	**GFDL_ESM2G**	**CanESM2**
3	INMCM4	CSIRO_MK3	**CanESM2**	**CCSM4**
4	IPSL-CM5A-LR	INMCM4	BCC_CSM1_1m	IPSL-CM5B-LR
5	**CanESM2**	IPSL-CM5A-LR	**IPSL-CM5A-MR**	MIROC5
6	**GFDL_ESM2G**	**CCSM4**	MRI-CGCM3	GFDL_ESM2M
7	BCC_CSM1_1	CNRM_CM5	CNRM_CM5	**IPSL-CM5A-MR**
8	GFDL_ESM2M	**CanESM2**	**CCSM4**	BCC_CSM1_1
9	IPSL-CM5B-LR	MRI-CGCM3	HadGEM2-CC	BCC_CSM1_1m
10	MIROC5	**GFDL_ESM2G**	IPSL-CM5A-LR	**GFDL_ESM2G**

773

774

775 **Figures**



776

777 Figure 1. Study Area, Columbia River Basin (CRB) in the Pacific North-West USA

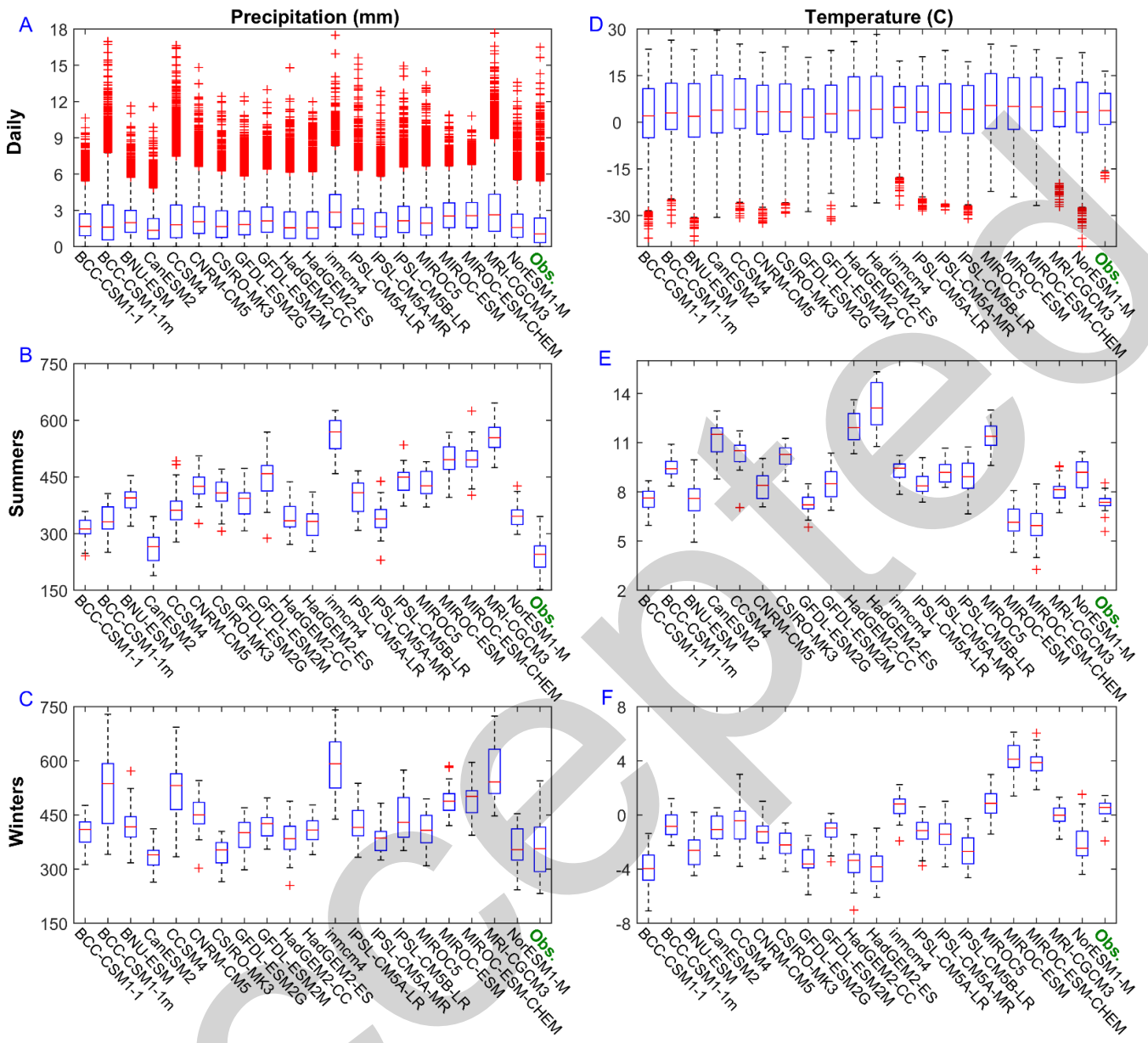


Figure 2. Boxplots depicting the distribution of precipitation and temperature in models and observation. Precipitation is plotted on the left, and temperature on the right. Daily, and seasonal data distribution are plotted from top to bottom, respectively. In each plot, observation is plotted after all GCMs, and is specified by green label.

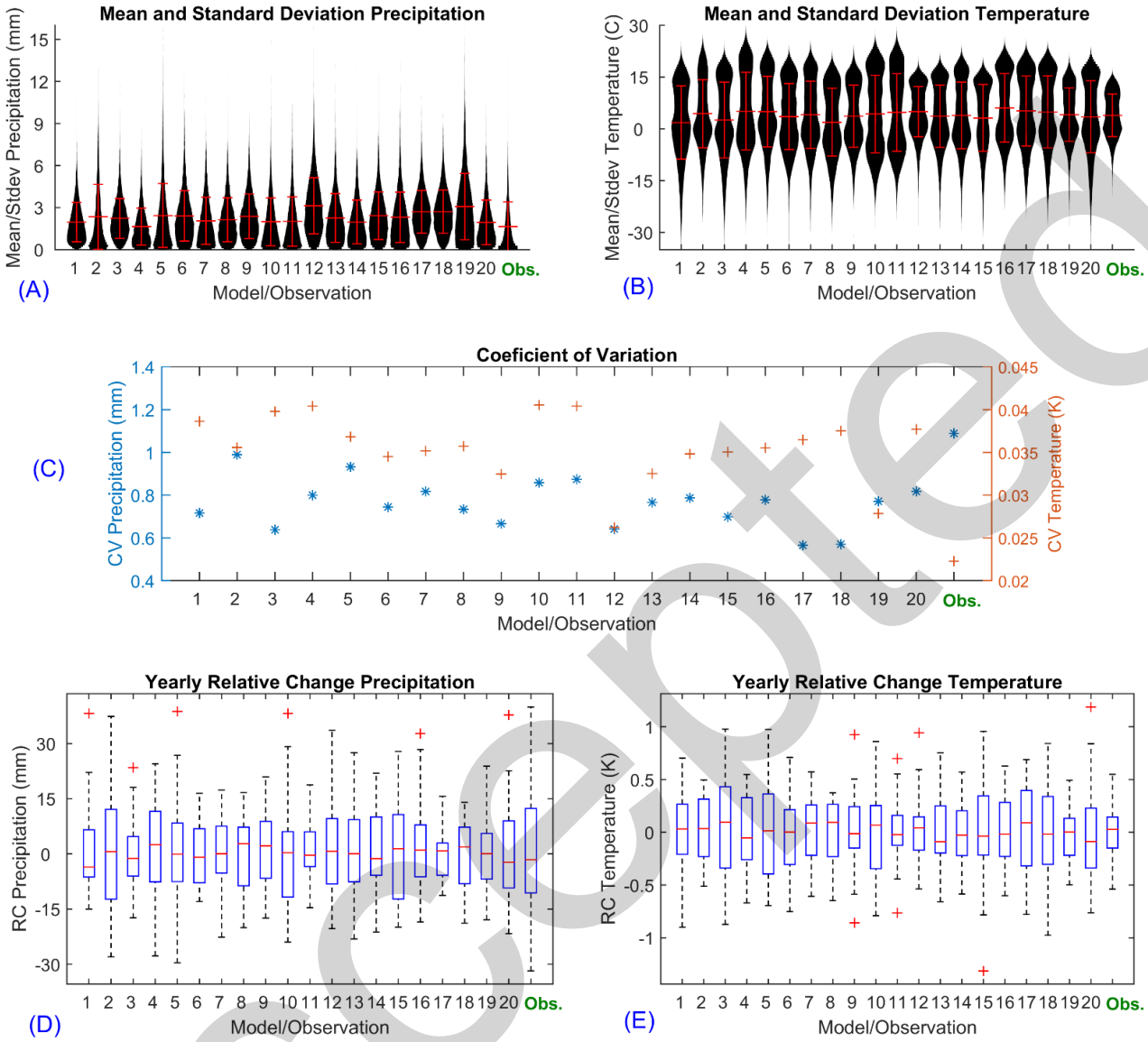


Figure 3. (a) Mean and (+/- 1) Standard Deviation of precipitation in models and observation (Violon Plot), (b) Mean and (+/- 1) Standard Deviation of temperature in models and observation (Violon Plot), (c) Values of CV for precipitation and temperature. Precipitation is depicted using '*' with values on the left y-axis; whereas temperature is depicted using '+' with values on the right y-axis, (d) Box plot of RC for precipitation in all the 30 years of data analysis, and (e) Box plot of RC for temperature in all the 30 years of data analysis. Model numbers on x-axis are the same as those provided in table 1.

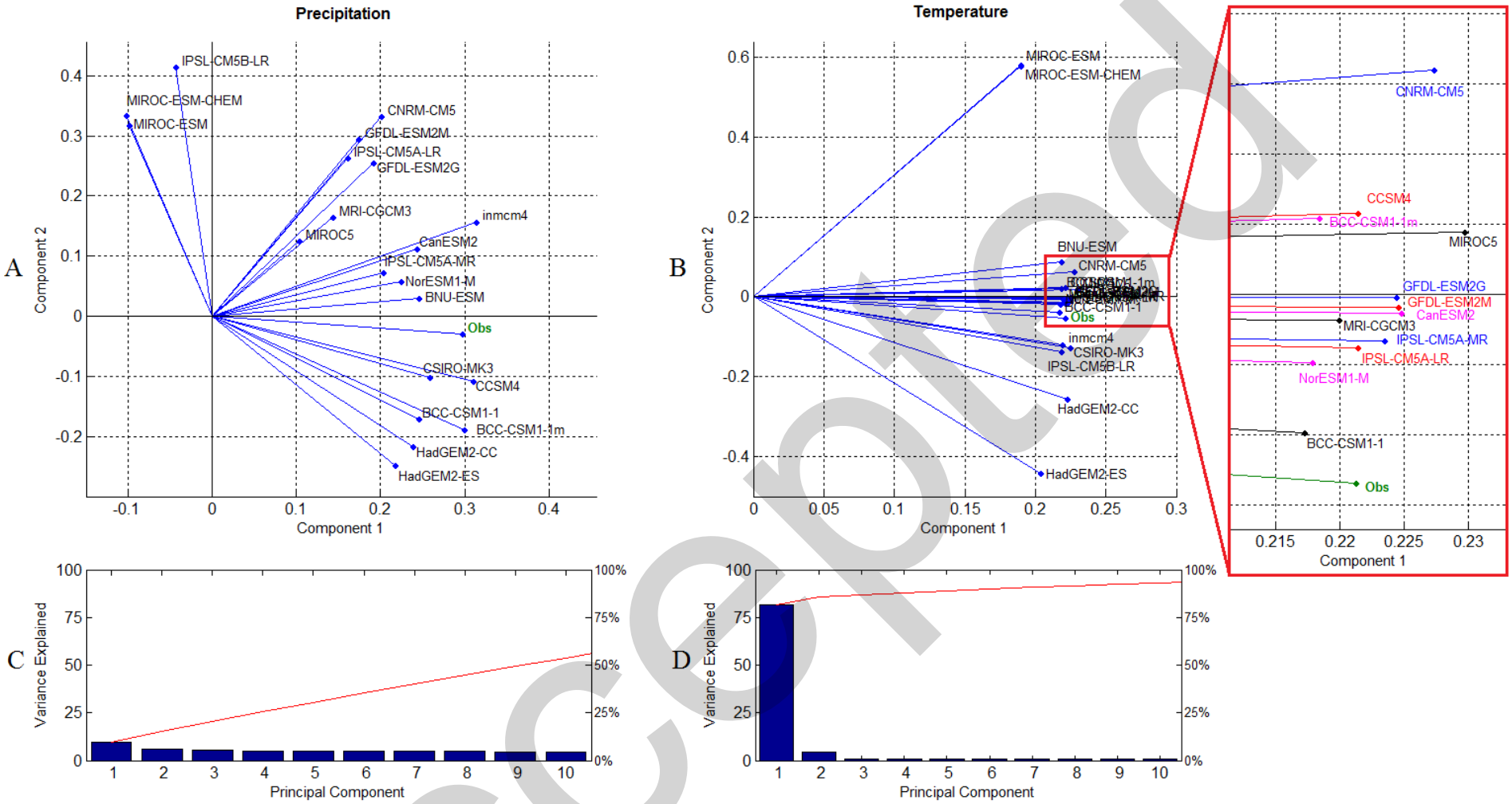


Figure 4. (a) Plots of relative distance of GCMs and gridded observational data on particular component of PCA for precipitation. Relative distance of GCMs on the axis of principal components compared to gridded observational data represents their proximity to observation, and is used to rank GCMs (the lower the distance, the closer the GCM predictions are to the gridded observational data), (b) Same as (a) for temperature, (c) Pareto plot (individual variance explained by principal components are represented in descending order by bars, and the cumulative total of variance is represented by the line) for total variance explained for a particular PCA component for precipitation, and (d) Same as (c) for temperature.

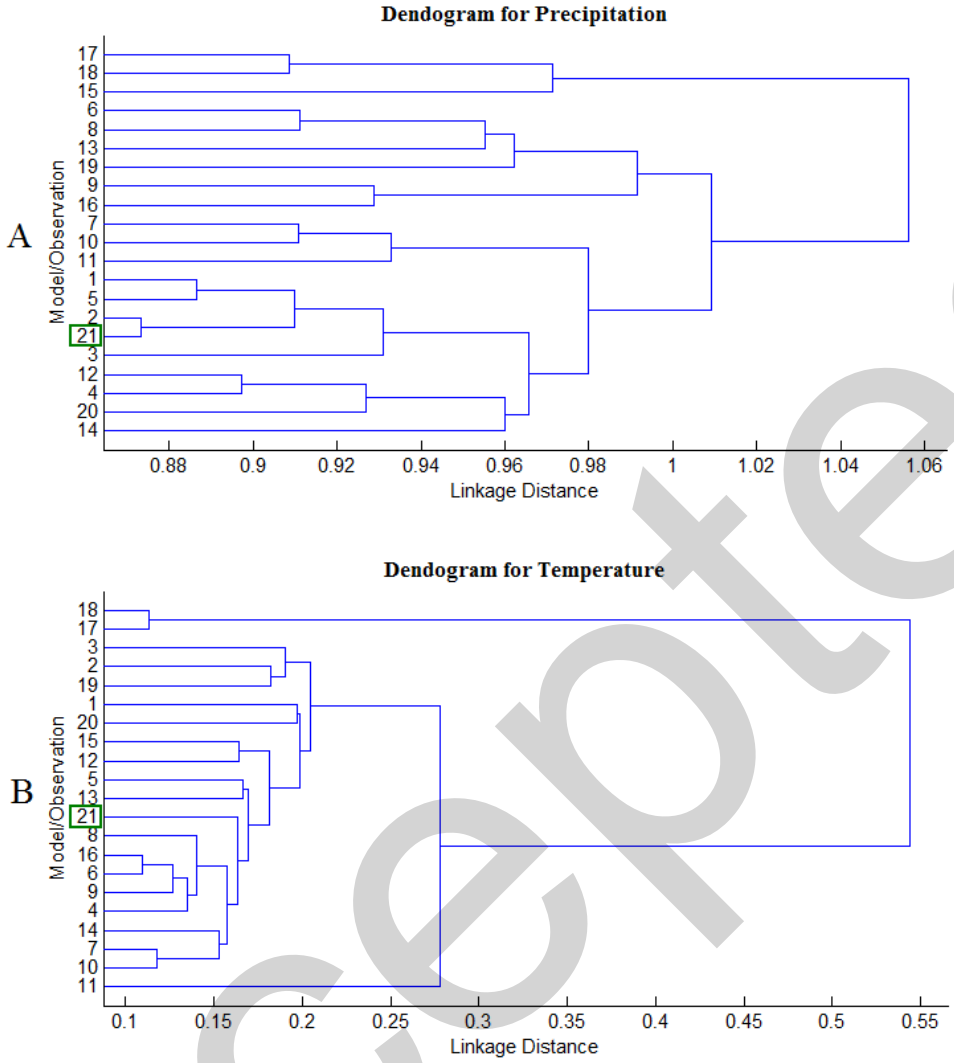


Figure 5. Dendograms generated with cluster analysis. (a) Cluster plot for precipitation dataset and (b) Cluster plots for temperature dataset. Linkage distance (between gridded observational data and GCMs) forms the basis of relative performance of GCM. Successive order of linkage is used to find the proximity of model to gridded observational data.

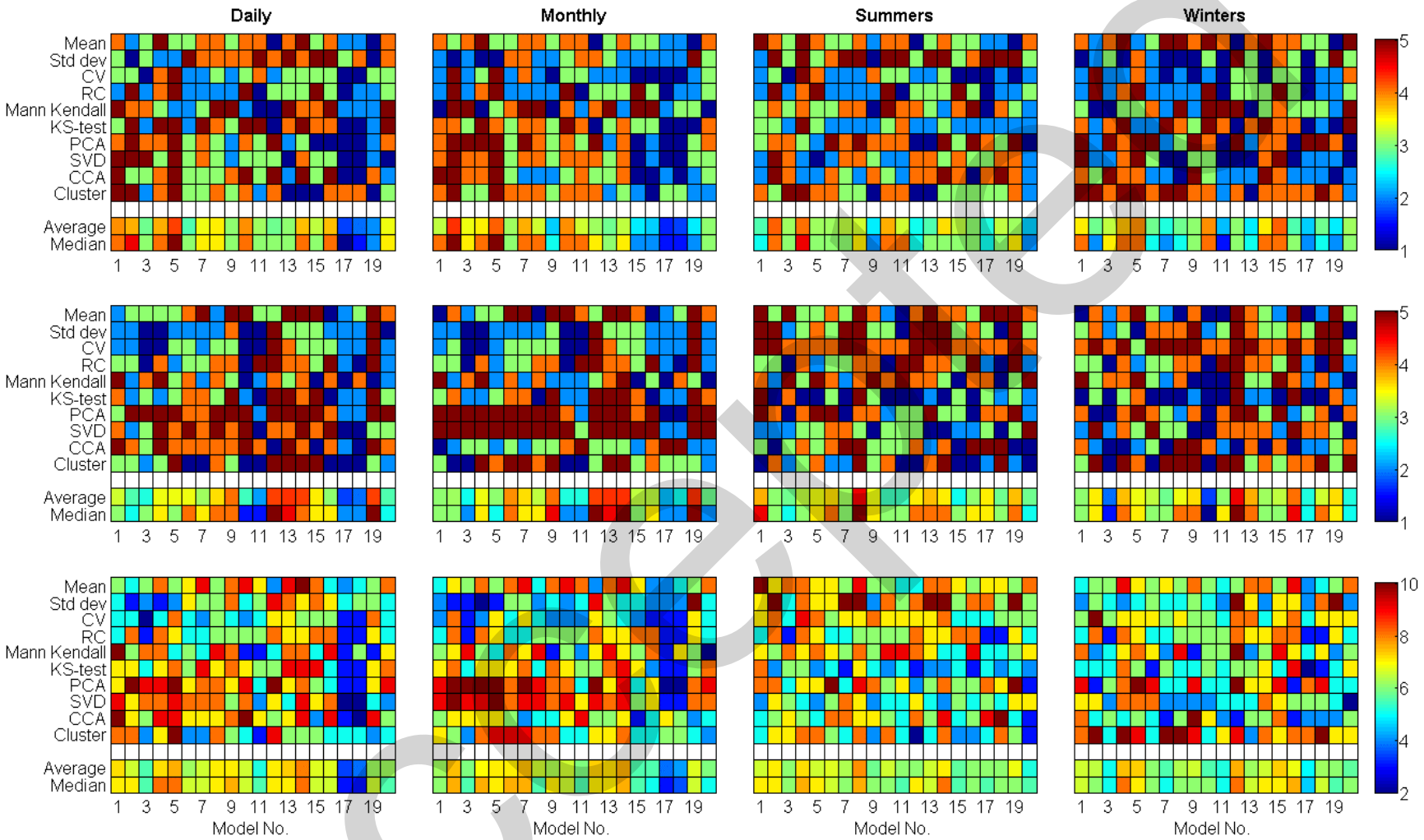


Figure 6. Performance of GCMs as evaluated against gridded observational dataset (Livneh et al. 2013) in each metric based on daily, monthly, and seasonal (summer and winter) data for precipitation (top), temperature (middle), and overall performance (bottom). In each plot, mean and median of all metrics are provided for each model in the last two rows.

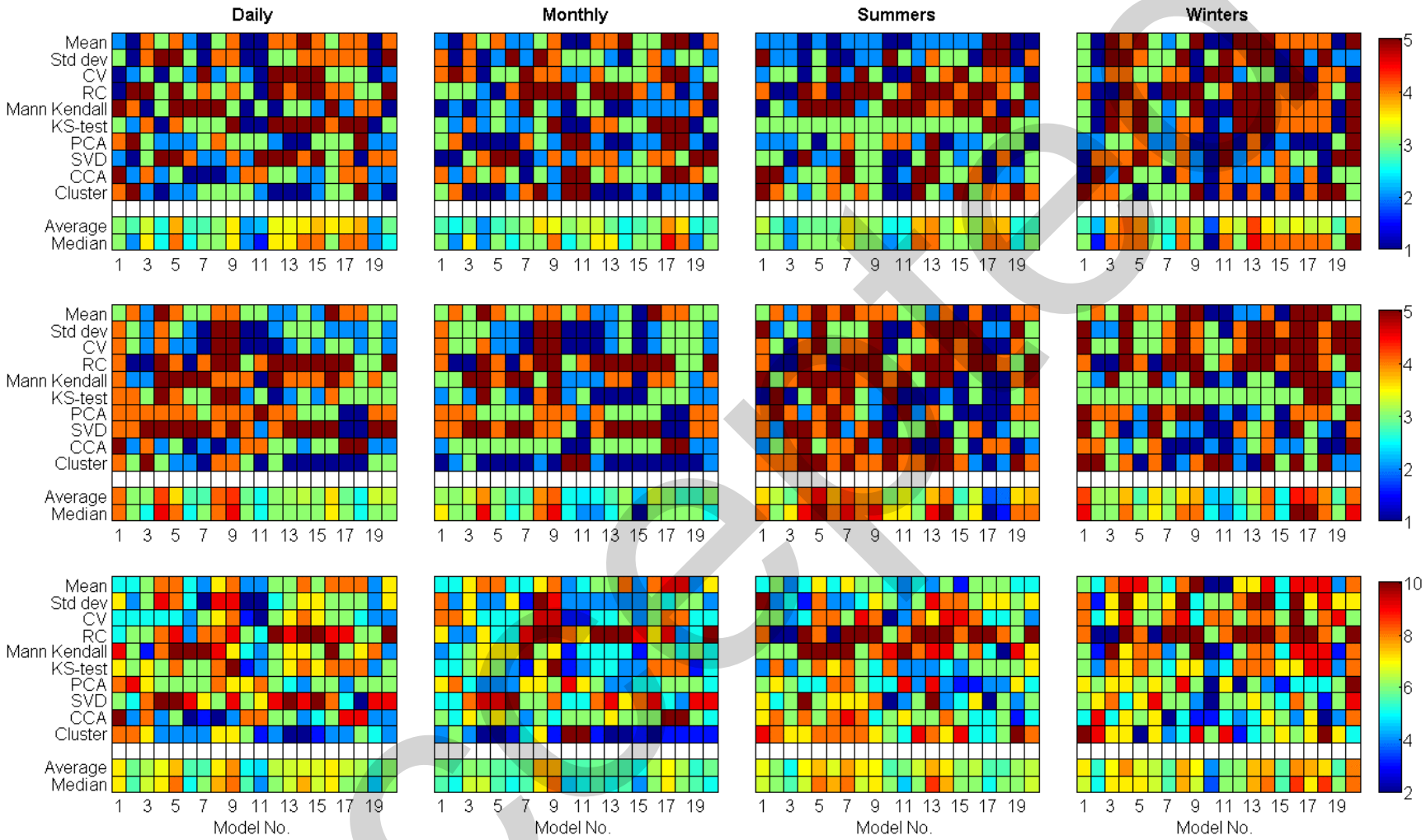


Figure 7. Performance of GCMs as evaluated against changed gridded observational dataset (Abatzoglou 2013) in each metric based on daily, monthly, and seasonal (summer and winter) data for precipitation (top), temperature (middle), and overall performance (bottom). In each plot, mean and median of all metrics are provided for each model in the last two rows.